**CW Assignment**

**Classification & Clustering**

This assignment is based on data analysis and machine learning approach to apply on diabetes dataset to predict the diabetes, such as K-Means clustering algorithm. There is two parts of this assignment one analysis with visualization and model training, the second part is model improvement that we are going to describe one by one.

**Part 1. Building up a basic predictive model**

1. **Data cleaning and transformation**

In this part we import the required libraries that is used such as NumPy, Pandas and Matplotlib.

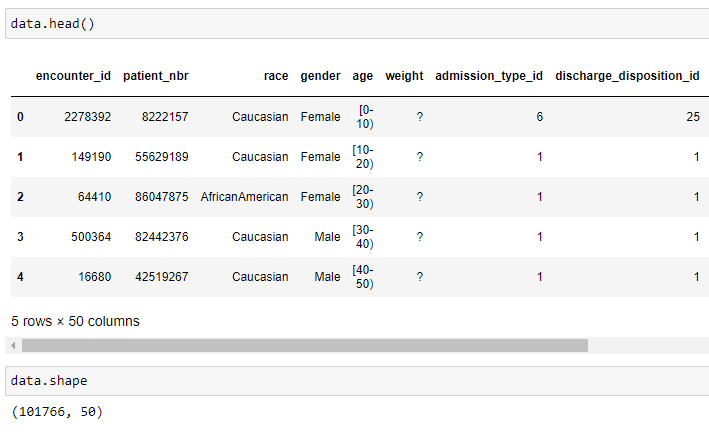


Figure 1

Next to read the dataset in csv format and check the head the of the dataset, then print the shape of the dataset that has 100k rows and 50 columns.

The for loop checks one by one all missing values such we find are replaced by np.nan values as the result can be seen in the figure.

The next method that checks the percentage of Nan values in columns and drop those Nan values and that drops the 2 columns that we can see. The perc to assign the 50% and calculate the minimum count and then check to drop all columns that have more than 50% missing values.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Figure 3

In the above figure we create a list execute the loop to check the columns that have 95% same values, if those columns have more than 95 percent and those columns are added to a list name “drop\_columns\_name”, and next we check the length of those columns that it shows 16 columns and those all are going to be dropped.

Graphical user interface, text, application, email

Description automatically generated

Figure 4

Here we can see first script prints the columns that have more 95% same values, the next script drops those columns permanently from dataset, and shape are changed from 47 to 31 columns.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Figure 5

In figure 5 we replace all Nan missing values with 0 in “diag\_1, diag\_2, and diag\_3” and drop all rows with missing values.

Categorical features are those that have object datatypes and numerical features are int64 and float64, but we distribute objects dtypes into categorical features and int64 dtypes into numerical and then print both.

Graphical user interface, text, application, Word

Description automatically generated

Figure 6

First script assigns the 3 standard deviations and checks quartiles, find the mean of columns, and divided by the standardized deviation columns with absolute values and then check the condition if it is less than 3.

1. **Data Exploration**

In this section we visualize the data according to the questions using plots, where visualization is the most important part for any king of data analysis work, it means it can be easily understandable by anyone.

First, we assign 0 to that those values were NO in “readmitted” column and assign 1 to those are greater/less than 30.

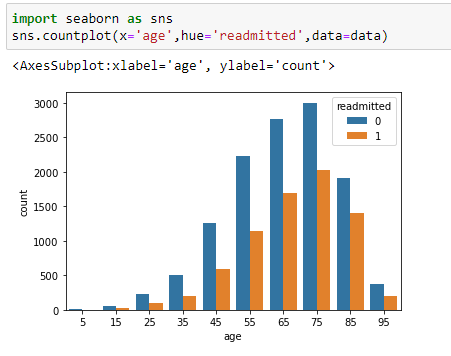


Figure 8

In this first script we import the seaborn library (a statistical library for data visualization) plot the “age” column data with x label, and “readmitted” with y label from our analysed dataset.

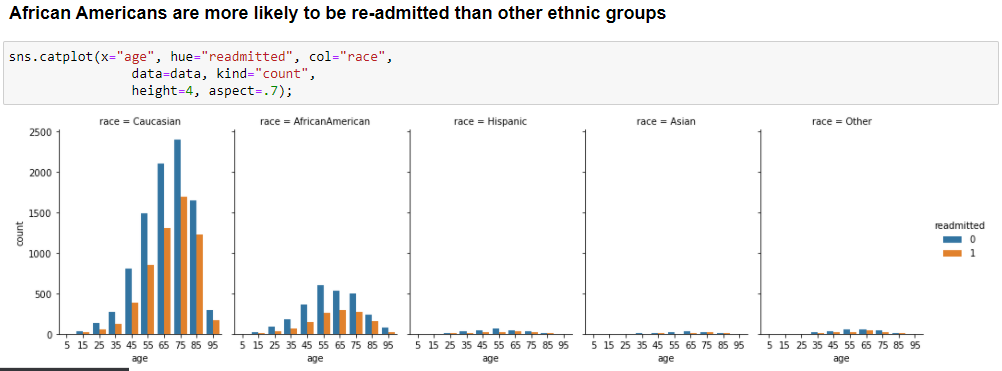
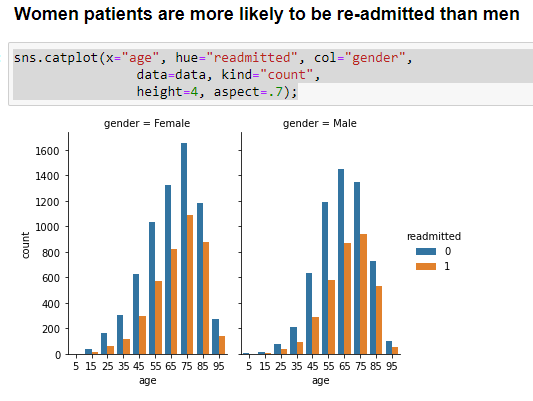


Figure 9

As here the catplot is used for better data visualization with multiple (age, readmitted, and race) columns that we can see overall ages of ethnic groups where Caucasian, African American, Hispanic, Asian, and Other are the races and their values are visible easily and readable.



Figure

Here as in this figure the female patients’ readmission values are more than male.

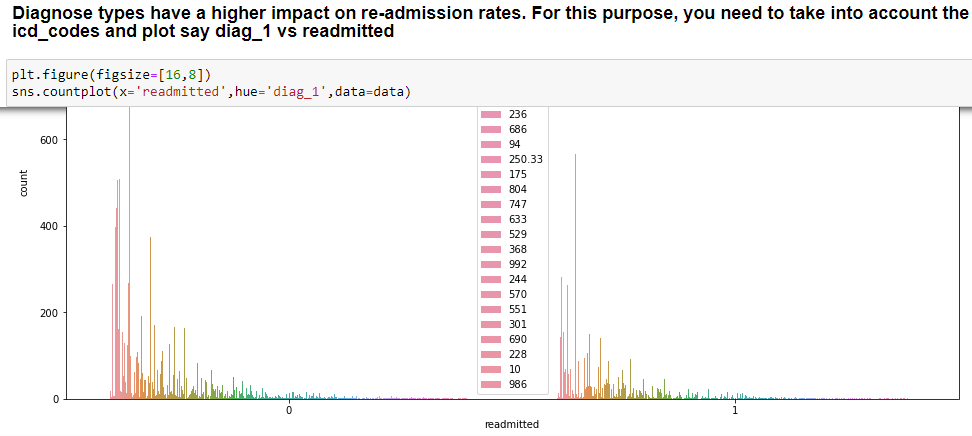


Figure 11

This plot shows lots of distinct values of diag\_1 with readmitted where the diag\_1 column’s value is hard to count.

1. **Model Building**

After the analysis and visualization processes next, we can build a model to predict the diabetes patients that they can be readmitted or not. Here we use the Random Forest classification model.

A **Random Forest** is an ensemble technique able to appear both regression and category duties with using multiple selection wood and a way known as Bootstrap and Aggregation, typically called bagging. The primary concept in the again of this is to mix a couple of selection bushes in figuring out the very last output as opposed to relying on individual decision bushes.

Random Forest has more than one preference bushes as base studying fashions. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for each version. This thing is known as Bootstrap. We want to approach the Random Forest regression method like some other device studying technique.

**Train and evaluate the model**

First, we separate the numerical columns from dataset and store in a list, then change the datatype of *readmitted* column from object to int., furthermore, split the data into training and testing that the 80% is training and 20% is testing.

Graphical user interface, text, application

Description automatically generated

Figure 12

Import the RandomForestClassifier model, to fit the x train and y train values, then predict the x test values as it shows the x and y test accuracy score. Next script we import the KFold model with validation accuracy score where K-Fold is a model that is used with cross\_val\_score library to find out the cross-validation score. Here we can see the average accuracy score is 64%.

Graphical user interface, text, application, email

Description automatically generated

Figure 13

We import the confusion matrix and print the report classification result and then sorted the “readmitted” column data and predict that data with confusion matrix method to plot the data heatmap graph.

Chart, treemap chart

Description automatically generated

Figure 14

Here is the result of confusion matrix and shows in a plot.

**Part 2. Improve Model**

**K-Means Clustering**

The k-means clustering method is an unsupervised machine mastering method used to understand clusters of facts gadgets in a dataset. There are many fantastic styles of clustering strategies, but okay-method is one of the oldest and maximum approachable. These tendencies make imposing appropriate sufficient-technique clustering in Python quite honest, even for beginner programmers and records scientists.

In this part we import the K-Means model from sklearn library and assign 2 number of clusters to fit the x train values and predict with x test as see the result of prediction in the form of 1 and 0’s.

Graphical user interface, text, application, email

Description automatically generated

Figure 15

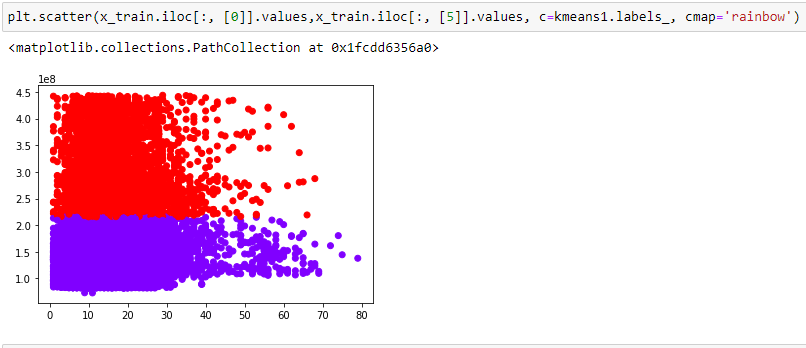


Figure 16

Here is the result of K-means model clustering of our data.